Contribution-based prioritization of LCI database improvements

Reinhard, Juergen; Wernet, Gregor; Zah, Rainer; Heijungs, Reinout; Hilty, Lorenz M.

published in
International Journal of Life Cycle Assessment
2019

DOI (link to publisher)
10.1007/s11367-019-01602-0

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Reinhard, J., Wernet, G., Zah, R., Heijungs, R., & Hilty, L. M. (2019). Contribution-based prioritization of LCI database improvements: the most important unit processes in ecoinvent. *International Journal of Life Cycle Assessment*, *24*, 1778-1792. https://doi.org/10.1007/s11367-019-01602-0

The International Journal of Life Cycle Assessment (2019) 24:1778–1792
https://doi.org/10.1007/s11367-019-01602-0

DATA AVAILABILITY, DATA QUALITY

Contribution-based prioritization of LCI database improvements: the most important unit processes in ecoinvent

Juergen Reinhard[1,2] · Gregor Wernet[3] · Rainer Zah[4] · Reinout Heijungs[5,6] · Lorenz M. Hilty[1,2]

Received: 22 August 2016 / Accepted: 22 February 2019 / Published online: 7 March 2019

Abstract

Purpose Improving the quality and quantity of unit process datasets in Life Cycle Inventory (LCI) databases affects every LCA they are used in. However, improvements in data quality and quantity are so far rather directed by the external supply of data and situation-driven requirements instead of systematic choices guided by structural dependencies in the data. Overall, the impact of current data updates on the quality of the LCI database remains unclear and maintenance efforts might be ineffective. This article analyzes how a contribution-based prioritization approach can direct LCI update efforts to datasets of key importance.

Methods A contribution-based prioritization method has been applied to version 3 of the ecoinvent database. We identified the relevance of unit processes on the basis of their relative contributions throughout each product system with respect to a broad range of Life Cycle Impact Assessment (LCIA) indicators. A novel ranking algorithm enabled the ranking of unit processes according to their impact on the LCIA results. Finally, we identified the most relevant unit processes for different sectors and geographies.

Results and discussion The study shows that a relatively large proportion of the overall database quality is dependent on a small set of key processes. Processes related to electricity generation, waste treatment activities, and energy carrier provision (petroleum and hard coal) consistently cause large environmental impacts on all product systems. Overall, 300 datasets are causing 60% of the environmental impacts across all LCIA indicators, while only 3 datasets are causing 11% of all climate change impacts. In addition, our analysis highlights the presence and importance of central hubs, i.e., sensitive intersections in the database network, whose modification can affect a large proportion of database quality.

Conclusions Our study suggests that contribution-based prioritization offers important insights into the systematic and effective improvement of LCI databases. The presented list of key processes in ecoinvent version 3.1 adds a new perspective to database improvements as it allows the allocation of available resources according to the structural dependencies in the data.

Keywords Contribution analysis · Life Cycle Assessment · Life Cycle Inventory database management · Meta-analysis · Prioritization

Responsible editor: Yi Yang

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11367-019-01602-0) contains supplementary material, which is available to authorized users.

✉ Juergen Reinhard
juergen.reinhard@quantis-intl.com; juergen.reinhard@empa.ch

[1] Department of Informatics, Informatics and Sustainability Research Group, University of Zurich, Binzmuehlestrasse 14, 8050 Zurich, Switzerland

[2] Empa Materials Science and Technology, Technology and Society Laboratory, Ueberlandstrasse 129, 8600 Duebendorf, Switzerland

[3] ecoinvent, Technoparkstrasse 1, 8057 Zürich, Switzerland

[4] Quantis Sarl, Konradstrasse 32, 8005 Zurich, Switzerland

[5] Institute of Environmental Sciences (CML), Leiden University, Einsteinweg 2, 2333 CC Leiden, The Netherlands

[6] Department of Econometrics and Operations Research, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

1 Introduction

Life Cycle Assessment is a technique for the comprehensive, quantitative assessment of the environmental impacts of products[1] throughout their entire life cycle (Finnveden et al. 2009). This involves the mapping of complex globalized networks consisting of thousands of interlinked human activities (Hellweg and Milà i Canals 2014). Tracing and measuring the exchange flows of and between these activities is realized on the basis of nodes called unit processes. A unit process represents one specific activity or a group of activities allocated to one unique output and records (i) the *exchanges with environment*, i.e., the input of natural resources and output of emissions; (ii) the *intermediate exchanges* from and to the technosphere, i.e., the input of usable energy and raw materials and the output of products and waste (Reinhard et al. 2016).

A product system represents the complete network of all unit processes involved in the life cycle or partial life cycle of a product (depending on the scope of the product under analysis). A typical product system covers thousands of unit processes, each of which needs to be described with exchange flow values (Bourgault et al. 2012). This information cannot usually be gathered as primary data within a specific project due to the high cost that would be involved in data collection (Reinhard et al. 2016). It is therefore a common practice to focus data collection efforts on selected activities that reflect the space for action—these activities are together called the foreground system—and to use generic data from Life Cycle Inventory (LCI) databases,[2] such as ecoinvent (Weidema et al. 2013), to model the remaining activities, called the background system (Tillman 2000; Bourgault et al. 2012; Reinhard et al. 2016). Even when 100 processes are modeled with primary data, the foreground system would still not exceed 5% of the entire product system as it typically involves thousands of unit processes (Steubing et al. 2016). Bearing this in mind, background data from LCI databases can be considered the backbone of any LCA study (Reinhard et al. 2016). The unit processes in such background databases therefore form the basic building blocks required by all LCA applications (Sonnemann and Vigon 2011).

However, unit process datasets are subject to uncertainties. The exchange flow data required for the accurate compilation of a unit process can be unavailable, wrong, or unreliable (Heijungs and Huijbregts 2004; Ciroth et al. 2016). As unit processes typically represent average conditions of a whole country, a given time period, and different instances of real processes, natural variability is always present (Huijbregts 1998). Both cases affect the overall accuracy of the unit process (Sonnemann and Vigon 2011). Consequently, LCI databases are under continuous extension and improvement. For example, ecoinvent has issued five updates in the past 5 years.

However, improvements are becoming progressively difficult due to the increasing numbers of datasets stored in existing databases. For example, version 3.5 of the ecoinvent database includes roughly 17,000 unit processes. If updating one unit process would, on average, require only one person-day,[3] a systematic update of the entire database within 1 year would require the continuous work of more than 75 researchers. Capacities for such extensive improvement efforts are typically not available and consequently LCI database improvements have to be prioritized.

To date, prioritization of LCI database improvements is mainly driven by the "continuing evolution in consumer preferences, and market and industry imperatives, and public policy" (Sonnemann and Vigon 2011). In addition, routines are in place to monitor existing datasets and to integrate newly available technological information on processes and product systems, on raw data, and on elementary flows (Sonnemann and Vigon 2011). Such improvements are important and can have a significant influence on the adequateness, representativeness, and actuality of LCA results. However, exclusively relying on this type of prioritization is "probably not the most effective use of resources in improving overall database quality" (Mutel 2012). Overall, the impact of current data improvements on the quality of the LCI database remains unclear and maintenance efforts might be ineffective.

We believe that insights into relative process relevance in existing LCI databases are one crucial—and currently missing—condition to organize data collection efforts more effectively. Knowledge about relative process relevance in a LCI database would offer a valuable perspective for "doing the right thing," namely improving the data elements with the largest influence on overall database quality. The goal of this article is therefore to rank the unit processes of ecoinvent 3.1 considering their importance across 19 selected LCIA indicators according to the following steps:

- We apply and improve the prioritization method from Reinhard et al. (2016) to the version 3.1 of the ecoinvent database. We first focus on three selected LCIA indicators

[1] The term product includes both goods and services.

[2] Our definition of LCI database follows the definition of the Shonan Guidance Principles (Sonnemann and Vigon 2011).

[3] The workload for updating a dataset varies greatly. Depending on the given level of completeness of a dataset, it might take weeks (e.g., when a global dataset is disaggregated into many country specific datasets) or only 1 h (e.g., when the quantity of one emission flow in one dataset is updated). Updating a dataset typically involves data collection, data entry/manipulation, and data submission to peer review (using a software tool called EcoEditor). Any change in an existing datasets or the submission of new datasets requires a peer review to be accepted into the database (Weidema et al. 2013). Therefore, one person-day should be considered as a rough but realistic estimate of the average effort associated with updating a dataset.

in order to establish a basic understanding and to highlight important characteristics of the approach.

- We identify and present the most relevant unit processes according to a set of 19 selected LCIA indicators using a newly developed ranking algorithm. The algorithm calculates the overall rank of unit processes considering their relevance across all LCIA indicators. We use the final prioritization list to identify the sectors and locations of particular relevance.
- We discuss and analyze possible reasons for the presence of systemic datasets and associated insights. We also define the limitations of the approach and related future work.

Section 2 presents the method and its implementation and Section 3 presents the results of applying the method to ecoinvent. Section 4 discusses these results and the merits of the method. Section 5 concludes with some final remarks.

2 Methods

2.1 Contribution-based prioritization

The application of contribution analysis (CA) is quite common in LCA and is implemented in all of the commercially available software tools (Goedkoop and Oele 2004; Ciroth et al. 2016; Ebner 2013). CA focuses on the disaggregation of aggregated results into a number of elements to identify the ones with the highest contributions (Heijungs and Kleijn 2001). Our elements of interest are the relative contributions of each unit process throughout each product system represented with respect to different LCIA[4] indicators.

Reinhard et al. (2016) have formalized CA for matrix-based LCI databases using the matrix inversion approach. Their method allows the database-wide computation of relative process contributions according to two perspectives:

- The *causer perspective (cau)* focuses exclusively on the elements of each unit process *directly causing* environmental interventions, i.e., the direct exchanges with the environment (resources consumed and emissions released) that cause environmental impacts. The database-wide application of the causer perspective helps to pinpoint the unit processes with consistently large contributions in terms of environmental interventions.
- The *connector perspective (con)* focuses exclusively on the *connecting elements* of each unit process, that is, the intermediate exchanges (e.g., the input of usable energy and raw materials) from other processes of the technosphere. The database-wide application of the connector perspective helps to pinpoint the unit processes that consistently *link* to large *upstream contributions*.

Reinhard et al. (2016) provide a comprehensive explanation of the method on the basis of a contrived example. In the following example, we therefore focus on the illustration of selected steps and results. SI3 (Electronic Supplementary Material) provides our implementation of the overall prioritization procedure in MATLAB for the illustrated example and the contrived exampled given in (Reinhard et al. 2016).

Figure 1 shows a streamlined workflow of the method on the basis of a very simple LCI database example.[5] Our example database is shown in matrix form and consists of three unit processes (X, Y, and Z), three corresponding products (x′, y′, and z′), and three elementary flows (CO_2, CH_4, and iron) (see technosphere matrix **A** and biosphere matrix **B**, step 1). In the technosphere matrix **A**, inputs and outputs are listed by row and unit processes by column. Positive numbers represent the production of outputs, while negative numbers represent the consumption of inputs. For example, unit process X consumes 0.75 units of product z′, 0.2 units of product y′, and produces one unit of product x′. We generally assume that the technosphere matrix is non-singular and square. The biosphere matrix **B** defines the environmental interventions (resources consumed and emissions released) per unit of process. For example, unit process Y and unit process Z both emit 1 kg of CO_2 emissions per unit of process.

We calculate the relative LCIA contributions of each unit process according to the *causer perspective* on the basis of an adapted version[6] of the LCA standard calculation procedure (Fig. 1, steps 1–2, Eqs. 1–5). By computing the inverse of the technosphere matrix and by multiplying it with the identity matrix,[7] **I**, we can calculate the amount needed of each process to satisfy a unit amount of product demand for all products in the database, i.e., the supply matrix **S** (step 1, Eq. 1). For example, the product system associated with the production of one unit of product x′ involves one unit of process X, 0.2 units of process Y, and 0.77 units of process Z. We then

[4] In principle, CA can also be applied on the inventory level (Heijungs and Kleijn 2001). We focus on environmental impacts, since we believe that a CA on the inventory level is of little practical interest in the context of prioritization of improvement efforts.

[5] Table S1 in SI1 (Electronic Supplementary Material) provides detailed information on all matrices and equations involved and SI2 (tab. "LCI database example") shows the example in excel.

[6] The key adaptations concern the diagonalization (expressed by the hat,) of vector **e** in Eq. 3, which facilitates an in-depth analysis of the contribution per process and the use of a demand matrix **I** (instead of a demand vector **f**) (Reinhard et al. 2016).

[7] In this context, the identity matrix explicitly introduces a unitary product demand (reference flow) for all product systems (Heijungs and Suh 2002, p. 85). Algebraically, this multiplication has no effect, and for numerical efficiency, it can be omitted.

Input data

Contribution analysis (IPCC2007, GWP100a)

Data consolidation

Equations

1

A: Technosphere matrix			
	X	Y	Z
x'	1	0	0
y'	-0.2	1	0
z'	-0.75	-0.1	1
B: Biosphere matrix			
CO_2	0	1	1
CH_4	0	0	0
Iron	-0.7	-1.2	-0.1

W: Characterization matrix			
LCIA-indicator	CO_2	CH_4	Iron
IPCC 2007 (GWP100a)	1	25	0
Metal depletion	0	0	-5E-8

S	x'	y'	z'
X	1.00	0.00	0.00
Y	0.20	1.00	0.00
Z	0.77	0.10	1.00

e	X	Y	Z
CO_2 eq.	0.00	1.00	1.00

M_{cau}	x'	y'	z'
X	0.00	0.00	0.00
Y	0.20	1.00	0.00
Z	0.77	0.10	1.00
t_{cau}	**0.97**	**1.10**	**1.00**

R_{cau}	x'	y'	z'
X	0.00	0.00	0.00
Y	0.21	0.91	0.00
Z	0.79	0.09	1.00
Sum	**1.00**	**1.00**	**1.00**

2 ***Causer perspective***

ID	$ß_{cau}$	Rank
X	0.00	3
Y	0.37	2
Z	0.63	1

4

t_n	
X	0.97
Y	1.10
Z	1.00

p	X	Y	Z
	1.00	0.09	0.00

M_{cum}	x'	y'	z'
X	0.97	0.00	0.00
Y	0.22	1.10	0.00
Z	0.77	0.10	1.00

M_{con}	x'	y'	z'
X	0.97	0.00	0.00
Y	0.02	0.10	0.00
Z	0.00	0.00	0.00

3 ***Connector perspective***

R_{con}	x'	y'	z'
X	1.00	0.00	0.00
Y	0.02	0.09	0.00
Z	0.00	0.00	0.00
Sum	**1.02**	**0.09**	**0.00**

ID	$ß_{con}$	Rank
X	0.33	1
Y	0.04	2
Z	0.00	3

1 $\mathbf{S} = \mathbf{A}^{-1}\mathbf{I}$

2 $\mathbf{e} = \mathbf{wB}$

3 $\mathbf{M}_{\text{cau}} = \hat{\mathbf{e}}\mathbf{S}$

4 $(\mathbf{t}_{\text{cau}})_j = \sum_{i=1}^{n} (\mathbf{M}_{\text{cau}})_{ij}$

5 $\mathbf{R}_{\text{cau}} = \mathbf{M}_{\text{cau}} \oslash (\vec{\mathbf{1}}\mathbf{t}_{\text{cau}})$

6 $(\boldsymbol{\beta}_{\text{cau}})_i = \frac{1}{n}\sum_{j=1}^{n} |(\mathbf{R}_{\text{cau}})_{ij}|$

7 $\mathbf{t}_{\text{n}} = \mathbf{t}_{\text{cau}}{}^{\text{T}} \oslash \hat{\mathbf{S}}$

8 $\mathbf{M}_{\text{cum}} = \hat{\mathbf{t}}_{\text{n}}\mathbf{S}$

9 $\mathbf{p} = \vec{\mathbf{1}}^{\text{T}} - (\mathbf{e} \oslash \mathbf{t}_{\text{cau}})$

10 $\mathbf{M}_{\text{con}} = \mathbf{M}_{\text{cum}} \circ (\vec{\mathbf{1}}\mathbf{p})^{\text{T}}$

11 $\mathbf{R}_{\text{con}} = \mathbf{M}_{\text{con}} \oslash (\vec{\mathbf{1}}\mathbf{t}_{\text{cau}})$

12 $(\boldsymbol{\beta}_{\text{con}})_i = \frac{1}{n}\sum_{j=1}^{n} |(\mathbf{R}_{\text{con}})_{ij}|$

Fig. 1 Workflow of the prioritization method explained on the basis of a very simple database example consisting of three processes (X, Y, and Z). The workflow describes the computation of one LCIA indicator (IPCC2007, GWP100a). The database is shown in matrix form (technosphere matrix **A** and a biosphere matrix **B**) with a corresponding characterization matrix **W** (step 1). We compute the relative contribution of all unit processes according to their causing (steps 1–2 and 4, Eqs. 1–6) and their connecting characteristics (step 1, 3 and 4, Eqs. 7–12) for the LCIA indicator IPCC2007, GWP100a, i.e., one row **w** of the characterization matrix **W**. The resulting matrices ($\mathbf{R}_{cau}$ and $\mathbf{R}_{con}$) express the relative contribution of a unit process (row) throughout the product system associated with the provision of product x′, y′, and z′ (column). We determine the average importance (rank) of a unit process based on the mean of absolute importance (β) of its contributions throughout all product systems (columns), e.g., for process Z in the causer perspective the mean of absolute importance is computed with $1/3 \times (0.79 + 0.09 + 1) = 0.63$ (step 4). Symbols used: the hat, "ˆ," indicates diagonalization of vector (Eqs. 3 and 7) OR filter-out off-diagonal elements of matrix (Eq. 7); "T" indicates transpose of matrix or vector; "⊘" in Eqs. 5, 7, 9, and 11 indicates element-by-element division; "∘" in Eq. 10 indicates element-by-element multiplication, i.e., the Hadamard product; $\vec{1}$ in Eqs. 5, 10, and 11 represents a vector of ones used to expand vector $\mathbf{t}_{cau}$ and **p** to matrix form. Table S1 in SI1 provides detailed information on all matrices and equations involved, SI2 (tab. "LCI database example") shows all matrices associated with the example in a more comprehensive manner and SI3 provides our implementation in MATLAB code

calculate the direct environmental impacts associated with the use of one unit of each process,[8] i.e., vector **e**, by multiplying a selected row (**w**) from the characterization matrix **W**—we use the first row, i.e., the LCIA indicator IPCC2007, GWP100a—with the biosphere matrix **B** (step 1, Eq. 2). Transforming the vector **e** into a diagonal matrix and multiplying it by the supply matrix **S** delivers the direct environmental impacts associated with every single process throughout all product systems, i.e., matrix $\mathbf{M}_{cau}$ (step 1, Eq. 3). The column-wise sum of $\mathbf{M}_{cau}$, $\mathbf{t}_{cau}$, then expresses the total environmental impact associated with the production of a unitary demand of each product in the database (step 1, Eq. 4). Finally, we compute, for each product system, the relative contribution of each process in relation to the contribution of all processes (step 2, Eq. 5) (Reinhard et al. 2016). The relative direct contribution matrix, $\mathbf{R}_{cau}$, expresses the relative process-specific contributions for each product, i.e., throughout the product systems associated with the provision of products x′, y′, and z′. For example, the unit processes Y and Z both have 1 kg of CO_2 emissions and, due to the characterization factor of one, also an impact equivalent to that of 1 kg of CO_2. Consequently, when involved in a product system, these unit processes have a relative contribution in the causer perspective. Vice versa, unit process X has no exchange of greenhouse gases and therefore has a direct contribution of zero throughout all product systems.

We calculate the relative contributions of each unit process according to the *connector perspective* (Fig. 1, steps 1 and 3, Eqs. 7–11) by first computing the cumulated (direct and upstream) environmental impacts of all processes throughout all product systems in the database (step 3, Eqs. 7[9] and 8) ($\mathbf{M}_{cum}$). We then calculate the relative proportion of each unit process which is caused upstream (**p**) and apply this proportion to the corresponding process contribution in $\mathbf{M}_{cum}$ to calculate $\mathbf{M}_{con}$, the process-specific upstream contribution throughout all product systems (step 3, Eqs. 9 and 10). We divide the process contributions in $\mathbf{M}_{con}$ by the corresponding total environmental impact per product, i.e., $\mathbf{t}_{cau}$, (step 3, Eq. 11). The resulting upstream contribution matrix, $\mathbf{R}_{con}$, expresses the relative process-specific upstream contributions associated with each product, i.e., throughout the product systems associated with the provision of products x′, y′, and z′. In this perspective, the

[8] Note that **e** expresses environmental impacts (in terms of CO_2 equivalents) and not the elementary flow CO_2.

[9] The operation in Eq. 7 ensures that the total environmental impacts refer to exactly one unit of a process and not one unit of product. It has no influence on our simple example but is important to achieve correct results in the presence of loops.

environmental impact of unit process X is fully determined by the environmental impacts associated with its intermediate exchanges, in this case the cumulated impacts associated with the provision of product y′ and product z′. Similarly, the environmental impacts of unit process Y are determined by the cumulated impact associated with the provision of product z ′. However, as process Y also has a direct emission, its relative upstream contribution is lower than the relative upstream contribution of process X. As unit process Z has no inputs of intermediate exchanges, it has a relative contribution of zero in the connector perspective.

Finally, we use the (arithmetic) mean of absolute importance[10] (β) to determine the average importance of the unit process throughout all product systems (columns) in the database (Fig. 1, step 4). In the causer perspective, the result, referred to as β_{cau} of a particular process, expresses the (absolute) mean contribution caused by its elementary flows throughout all product systems in the database (Eq. 6). That is, highly ranked unit processes in the causer perspective are important to the overall environmental impacts throughout the database because of their elementary flows. Highly ranked unit processes in the connector perspective are important because of their intermediate exchanges. For the connector perspective β_{con} expresses the (absolute) mean contribution which are transmitted by all intermediate inputs of a unit process throughout all product systems in the database (Eq. 12). To put it differently, it expresses the average loss throughout all product systems that would occur if we removed all intermediate inputs from a unit process. This characteristic results in double counting when adding up the individual elements in β_{con} for all processes in the connector perspective since the β_{con} of a particular process (double) counts (at least part of) the contribution already accounted for in the β_{con} of its connected processes. Therefore, and in contrast to β_{cau}, the sum of β_{con} can be much larger than one.[11] In fact, any cumulated sum of β_{con} must be interpreted as a theoretical maximum that expresses the amount of contribution which, on average, "flows through" each of the connectors in the database when they are measured independently of each other (Reinhard et al. 2016).

We sort the elements of vector β in descending order (see Eq. 13 and Fig. 2, step 1) and store the results into β_{sort}. This sorted vector is used to compute the cumulative contribution, hold by the h processes with the largest contribution. $(\beta_{sort\_cum})_h$ indicates the *cumulated mean of absolute importance* of the h processes with the largest contribution (see Fig. 2, step 1).

$$\beta_{sort}=\left(\beta_{(n)},\ \beta_{(n-1)},\ldots,\beta_{(1)}\right) \text{ where } (\beta_{sort})_i \le (\beta_{sort})_j \text{for all } i<j \tag{13}$$

$$(\beta_{sort\_cum})_h = \sum_{i=1}^{h} (\beta_{sort})_i \text{ for } h=1,\ldots,n \tag{14}$$

In order to express the mean of absolute importance in relation to the cumulated importance of all other processes, we compute the *mean of relative importance*, $(\beta_{sort\_rel})_i$, by dividing the $(\beta_{sort})_i$ of a particular process i by the sum of (β) all processes (Eq. 15).

$$(\beta_{sort\_rel})_i=\frac{(\beta_{sort})_i}{\sum_i^n \beta_i} \tag{15}$$

Finally, we need the cumulative relative importance, $\beta_{sort\_cum\_rel}$, defined by

$$(\beta_{sort\_cum\_rel})_h = \frac{(\beta_{sort\_cum})_h}{\sum_{i=1}^{n} \beta_i} \tag{16}$$

This workflow has to be executed separately[12] for every LCIA indicator, so for every row of **W**. It is applicable to any disaggregated LCI database that provides, or can be transformed into, the technosphere matrix **A** and a biosphere matrix **B**. Moreover, a corresponding characterization matrix **W** is needed.

2.2 Prioritizing across many LCIA indicators

Unit process importance refers to the ranking of one particular LCIA indicator, i.e., is determined by the process order in β_{sort}. Yet, process importance varies across LCIA indicators (Reinhard et al. 2016). That is, different unit processes are important across different LCIA indicators. Therefore, "robust prioritization requires computation across a large set of LCIA indicators" (Reinhard et al. 2016).

We aim to establish a ranking of unit *process relevance*, i.e., a ranking of unit processes considering their importance across *many* LCIA indicators. However, assessing the variable importance and establishing a corresponding ranking of process relevance across a large set of LCIA indicators is a resource intensive and a challenging task which is very sensitive to potential size differences[13] in the mean of absolute importance. To date, there is no algorithm which facilitates ordering of the unit processes in an LCI database according to their process relevance.

We developed an algorithm that uses the properties of the Lorenz curve to rank unit processes according to their relevance across any set of LCIA indicators. The Lorenz curve is

[10] We use the modulus for the calculation of β. This reduces distortions due to negative contributions which can result from negative characterization factors and/or negative elementary flows (e.g., an uptake of heavy metals by plants in a cultivation dataset) in the biosphere matrix.

[11] To date, a value lower than one was only observed for very simplistic database example (e.g., the sum of β_{con} is lower than one (0.36) in our very simplistic database example) but not for actual LCI databases.

[12] Note that, once computed, **S** can be reused without further adaptation and consequently one matrix inversion is sufficient for the computation of any set of LCIA indicators.

[13] As shown by Reinhard et al. (2016), the importance of one and the same unit process can vary a lot across different LCIA indicators.

1 q = IPCC 2007 (GWP100a)

ID	β_{sort}	β_{sort_cum}
Z	0.63	0.63
Y	0.37	1.00
X	0	1.00

1 q = Metal depletion

ID	β_{sort}	β_{sort_cum}
Y	0.41	0.41
Z	0.36	0.77
X	0.23	1.00

2 q = IPCC 2007 (GWP100a)

ID	h	$\beta_{sort\_cum\_rel}$
	0	0
Z	1	0.63
Y	2	1.00
X	3	1.00

Lower than ccv

Equal to ccv

Successor to ccv

2 q = Metal depletion

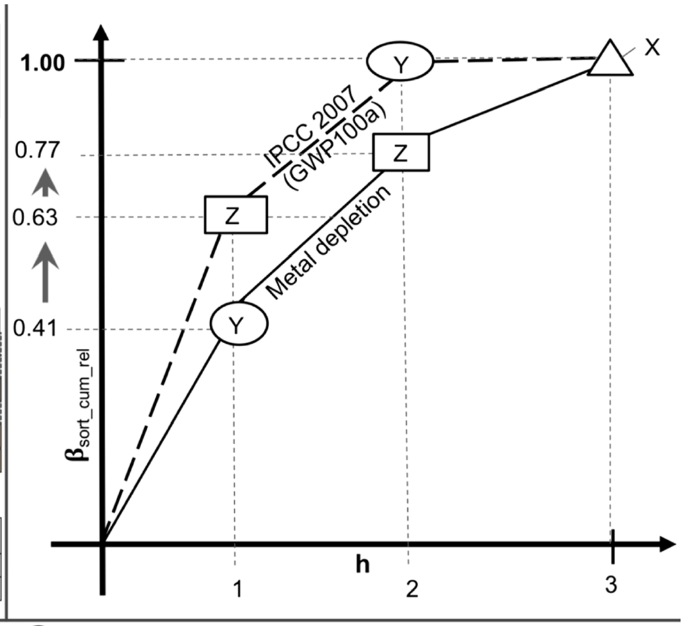
ID	h	$\beta_{sort\_cum\_rel}$
	0	0
Y	1	0.41
Z	2	0.77
X	3	1.00

3

q	ccv = 0.41 ID	h	$\beta_{sort\_cum\_rel}$	ccv = 0.63 ID	h	$\beta_{sort\_cum\_rel}$	ccv = 0.77 ID	h	$\beta_{sort\_cum\_rel}$	ccv = 1 ID	h	$\beta_{sort\_cum\_rel}$
q = Metal depletion	Y	1	0.41	Y	1	0.41	Y	1	0.41	Y	1	0.41
	Z	2	0.77	Z	2	0.77	Z	2	0.77	Z	2	0.77
	X	3	1.00	X	3	1.00	X	3	1.00	X	3	1.00
q = IPCC 2007	Z	1	0.63	Z	1	0.63	Z	1	0.63	Z	1	0.63
	Y	2	1.00	Y	2	1.00	Y	2	1.00	Y	2	1.00
	X	3	1.00	X	3	1.00	X	3	1.00	X	3	1.00

4 Prioritization table

LCIA indicator	0.41	0.63	0.77	1.00
Metal depletion (MDP)	Y	Y, Z	Y, Z	Y, Z, X
IPCC 2007 (GWP100a)	Z	Z	Z, Y	Z, Y, X



5

Process IDs	Cumulated contribution value			
	0.41	0.63	0.77	1.00
Y	1	1	2	2
Z	1	2	2	2
X				2

6

Rank	Process IDs (sorted)	Cumulated contribution value			
		0.41	0.63	0.77	1.00
1	Z	1	2	2	2
2	Y	1	1	2	2
3	X				2

Fig. 2 Ranking procedure, with illustrative numbers for the causer perspective. (Step 1) Note that column "β_{sort_cum}" results from sorting and cumulating β_{cau} (obtained from Fig. 1) in descending order. (Step 2) The value pairs in columns h and "$\beta_{sort\_cum\_rel}$" (left) are used to construct the MLCs (right). (Step 3) Using the MLC tables as a basis, we identify the process indices needed to equal or to succeed a given cumulated contribution value (ccv) for both LCIA indicators and pass them into a prioritization table (step 4) where they are stored in distinct reference to the corresponding ccv and the LCIA indicator. (Step 5) The unsorted ranking table indicates the cumulative process relevance, i.e., the importance of a particular process across all LCIA indicators for a particular ccv. (Step 5) The final ranking table is sorted in descending order, first according to the ccv in the first bin, then according to the ccv in the second bin, and so forth. SI3 provides our implementation of this procedure in MATLAB code

well known for comparing income and wealth inequality in the field of economics and provides the basis for statements such as "the bottom 50% of the households have only 10% of the total income" (Duclos and Araar 2006). We first construct a mirror image (inverse) of the Lorenz curve (MLC) for each LCIA indicator and perspective. That is, instead of plotting the lowest contribution first, we start with the process whose mean of absolute importance throughout the database is the largest and proceed by adding all other contributions in descending order. Our MLC indicates the cumulated mean of relative importance, held by h processes.

More formally, the discrete MLC is defined as the polygon segment joining the points $(h, (\beta_{sort\_cum\_rel})_h)$ for all $h = 1, 2, \ldots, n$ as well as the points (0, 0) and (n, 0).

Figure 2 shows the MLC tables (left) and corresponding polygon segments (right) of the example system shown in Fig. 1, for the impact categories IPCC 2007 (GWP100a) and metal depletion. It illustrates the general ranking procedure for the causer perspective. However, there it works in exactly the same way for the connector perspective. $\beta_{sort\_cum\_rel}$ represents the *sorted and cumulated mean of relative importance*, for each LCIA indicator q, with $q = 1, 2\ldots$ indicating the particular MLC table or LCIA indicator (see Fig. 2, step 2). Together with the unique IDs for each unit process, the data table for each MLC consists of three columns (see Fig. 2, step 2).

Our ranking algorithm operates on the basis of the zero-deprived[14] MLC tables (see Fig. 2, step 3). We start with the smallest cumulative relative contribution value (ccv) in the MLC tables and increase until 1. At every unique ccv—in our example case 0.41, 0.63, 0.77 and 1—the algorithm identifies, separately for each LCIA indicator, the process index/indices of the cumulated process contribution(s) which

1. Are *lower than* a given ccv
2. *Is/are equal to* a given ccv and

[14] As the zeros in the MLC tables do not relate to any process index, they are not required for the ranking algorithm.

3. Directly *follows/succeeds* a given ccv[15]

Note that conditions 2 and 3 are mutually exclusive. That is, if the algorithm identifies, in one MLC table, one or several cumulated process contribution(s) which *equal* a given ccv, then the third query is not executed for that MLC table and the algorithm moves to the next MLC table.

At each match, the algorithm passes[16] the corresponding process index/indices into a prioritization table where they are recorded in distinct reference to their ccv and the LCIA indicator (see Fig. 2, step 4). The column-specific process indices in the prioritization table facilitate statements such as "to equal or succeed a ccv of x% across *all* considered LCIA methods, the process/es with index/indices *t1, t2,* etc. is/are required." Figure 2, step 3, illustrates the core steps of the procedure by highlighting for each unique ccv the corresponding cumulated process contributions which are

- Lower than (gray rectangle, dashed line),
- Equal to (orange rectangle, straight line), or
- Which directly succeed the ccv (green rectangle, dotted line) across all MLC tables.

For example, the algorithm identifies process Y to equal a cumulated contribution of 0.41 for Metal depletion and—as we do not have a lower or an equal match in the IPCC2007 MLC table—the cumulative contribution which follows/succeeds the given ccv, namely process Z. Next, to equal or succeed a ccv of 0.63, the cumulative contribution of process Z and process Y (IPCC2007) and process Z (metal depletion) are required. To equal or succeed a ccv of 0.77, we require the cumulative contribution of process Y and Z (metal depletion) and process Z and Y (IPCC2007). Finally, to equal or succeed a ccv of 1, all processes are required.

In the next step, we compute the process frequency in distinct reference to the process ID (see Fig. 2, step 5), basically by counting for each process (row) and ccv (column) the process occurrence in the prioritization table (step 4) across all LCIA indicators,[17] i.e., by merging the process IDs prioritized by different LCIA indicators. The number in the ranking table indicates, for each process ID, the frequency this particular process is required to equal or succeed a given ccv across all considered LCIA indicators.

Finally, the algorithm ranks unit processes according to their frequency (Fig. 2, step 5) by sorting them in descending order, at first according to the lowest ccv (highest priority), then according to the second lowest ccv (2nd highest priority), and so forth, until the numbers across all ccv's are sorted. This ensures that the order of processes resembles their actual frequency according to their size classes and facilitates their ranking across any set of LCIA indicators. We will explore distinct properties of this algorithm in the discussion section.

2.3 Implementation

Version 3.1 of the ecoinvent database offers three system models. A system model consists of a predefined set of rules for the transformation of unlinked multi-output activities into interlinked, single-product processes. We work with the system model "Allocation, cut-off by classification" (cut-off).[18] The cut-off system model results in 11,304 unit processes, i.e., a square technosphere matrix of the size 11,304 × 11,304 (see Table S3 in SI1, Electronic Supplementary Material). All unit processes are allocated into single-product processes according to the cut-off approach.[19]

We first loaded the linked database model and 19 preselected LCIA indicators into MATLAB and applied Eqs. 1–5 and 7–11 (see Fig. 1) to compute $\mathbf{R}_{cau}$ and $\mathbf{R}_{con}$repeatedly for each LCIA indicator. We selected LCIA indicators based on their scientific quality[20] and their availability.

Next, we calculated the β of each process listed in $\mathbf{R}_{cau}$ and $\mathbf{R}_{con}$ (see step 6 and 7 in Fig. 1) according to all LCIA indicators.

Subsequently, we ranked the processes according to their mean of absolute importance (β) in descending order (β_{sort}) and compute β_{sort_cum}, β_{sort_rel} and the discrete MLC—for each LCIA indicator—defined by the polygon joining the value pairs in **h** and the corresponding cumulated relative contribution in $\beta_{sort\_cum\_rel}$. Due to the limitations in space, LCIA indicator-specific results are only presented for the indicators CC, Etox, and ReCiPe. We aimed to highlight important characteristics of the most important processes according to these different LCIA indicators.

Finally, we use the novel ranking algorithm to calculate an overall rank of unit processes considering their relevance across all 19 LCIA indicators. For the sake of efficiency, instead of using all $\beta_{sort\_cum\_rel}$ values across the 19 LCIA

[15] In the presence of more than one successor of equal size—as with 1.00 (Y) and 1.00 (X) in IPCC 2007 for a ccv of 0.77—the successor with the lower h index is selected.

[16] Only the index/indices which is/are not already recorded will be added to the LCIA indicator specific row of the prioritization table.

[17] One possibility is to vertically concatenate the rows (LCIA indicators) of the prioritization table (Fig. 2, step 3) into one matrix and then count the occurrence of each process ID separately for each ccv (column). SI3 (Electronic Supplementary Material) shows our implementation in MATLAB (see function "determineProcessrelevance").

[18] The system model is based on the cut-off approach where primary (first) production of materials is always allocated to the primary user of a material. Furthermore, a primary producer of a recyclable material does not receive any credit for its provision. Therefore, recyclable materials are available burden-free to recycling processes, and secondary (recycled) materials bear only the impacts of the recycling processes (Wernet et al. 2016).

[19] In this system model, processes are only allowed to have more than one output if the non-unitary output represents a recyclable product.

[20] Scientific quality is determined according to the recommendations in the ILCD-Handbook (EC-JRC 2011).

indicators, we work with a predefined set of 9 ccv, i.e., ccv = 0.1, 0.2, … , 0.9, which uniformly divides the value domain of $\beta_{sort\_cum\_rel}$. The complete prioritization list presented in SI2 (Electronic Supplementary Material), however, uses a more fine-grained division of 99 ccv.

3 Results

3.1 Prioritization according to selected LCIA indicators

Figure 3 visualizes the MLC's of the three selected LCIA indicators for the causer (left) and the connector (right) perspective. To improve readability, we used a logarithmic scale for the number of processes **h** (X-axis).

Tables 1 and 2 show the corresponding seven unit processes with the largest mean of absolute importance β, once for the causer and once for the connector perspective, respectively. The results are sorted in descending order according to the value of β. For the connector perspective, we also show—in addition to the mean of absolute process importance—the cumulated mean of absolute importance, $(\beta_{con})_{sort\_cum}$, and the mean of relative importance, $(\beta_{con})_{sort\_rel}$. While $(\beta_{con})_{sort}$ and $(\beta_{con})_{sort\_cum}$ provide insight into the absolute "throughput" associated with the intermediate flows of the most important unit processes, $(\beta_{con})_{sort\_rel}$ highlights the relative importance of a process contribution in relation to the mean contribution of all other connectors in the database. The full list covering all processes and all LCIA indicators is available in the SI2 (Electronic Supplementary Material).

Figure 3 in combination with Tables 1 and 2 reveals a remarkable concentration of process importance. In regard to the causer perspective, the seven processes with the largest relative importance already accumulate 21% (CC), 16% (ReCiPe), and 66% (Etox) (see column "$(\beta_{cau})_{sort\_cum\_rel}$" in Table 1). In other words, less than *one per mill* of the processes in the database already cause a significant proportion of the overall contribution throughout all product systems.

The seven processes with the largest relative importance in the connector perspective accumulate 5% (ReCiPe), 7% (CC), and 17% (Etox), respectively (see column "$(\beta_{con})_{sort\_cum\_rel}$" in Table 2 and Fig. 3). However, in terms of the absolute "throughput," the same processes accumulate 49% (CC), 42% (ReCiPE), and 169% (Etox) of the contribution (see column "$(\beta_{con})_{sort\_cum}$" in Table 2). For example, removing all intermediate flows from the process "sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site" would reduce environmental impacts of Etox throughout all product systems on average by roughly 42%. This highlights that the connecting elements (intermediate exchanges) of a process typically transmit more environmental impact (contribution) than caused by its causing elements (elementary exchanges). That is, the accuracy of the intermediate exchanges of the major connectors has a very large influence on the quality and the results of many product systems in the database.

Third, we can notice some similarities between the unit processes prioritized by the causer and the connector perspective, i.e., the same unit processes can be important according to both perspectives. For example, "electricity, high voltage//[CN] electricity production, hard coal" and "hard coal//[CN] hard coal mine operation" are consistently important according to both perspectives. At the same time, however, the most important connector, "petroleum//[GLO] market for petroleum" has—according to ReCiPe—no contribution in the causer perspective. The same holds true for many other connectors and becomes apparent when looking at the correlation between the two perspectives shown as a scatter plot (see Fig. S7 in SI1, Electronic Supplementary Material).

Fourth, we can spot some similarities across LCIA indicators, i.e., different LCIA indicators point to the same unit

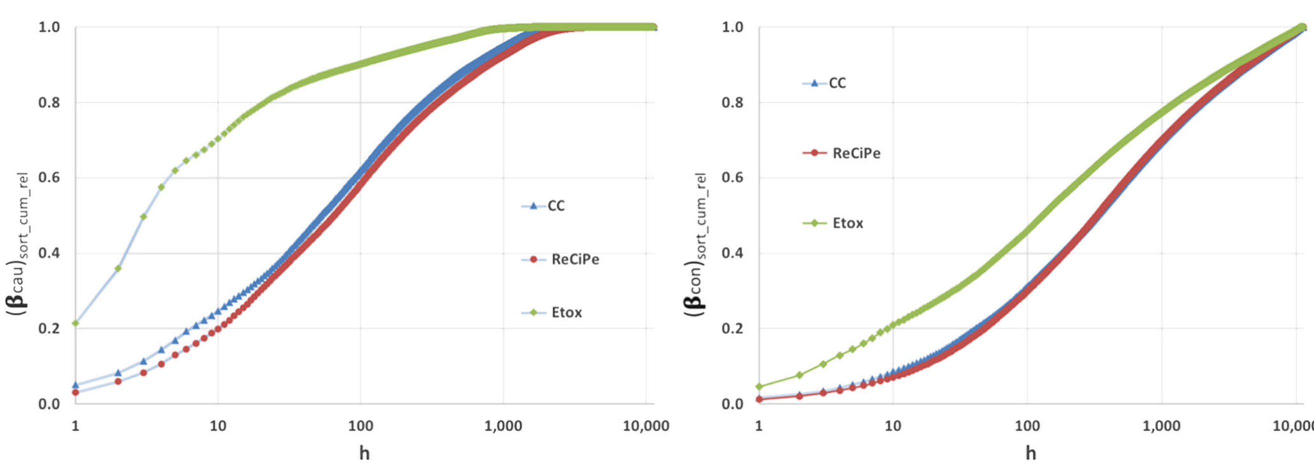


Fig. 3 MLCs for the causer (left) and the connector perspective (right) illustrating the cumulated mean of relative importance associated with **h** (1, 2,… 11,304) unit processes for three selected LCIA indicators

Table 1 Excerpt of the summary table for the seven causers with the largest mean contribution for three selected LCIA indicators. The cumulated mean of relative importance $(\beta_{cau})_{sort\_cum\_rel}$ of the seven causers with the largest mean contribution is highlighted in italic format. The full list can be viewed in the SI2. Ecoinvent uses geographical shortcuts (also shown in SI2): CN, China; RoW, Rest-of-the-World; GLO, global; RME, Middle East; RNA, northern America. RoW has been introduced to cover all locations where a local process is not yet available. Consequently, the spatial scope of RoW varies. Currently, it can represent more than 100 different locations (Wernet et al. 2016)

LCIA	Rank	Name [product //[geographical location] activity]	$(\beta_{cau})_{sort}$	$(\beta_{cau})_{sort\_cum\_rel}$
CC	1	Electricity, high voltage//[CN] electricity production, hard coal	0.049	0.05
	2	Hard coal//[CN] hard coal mine operation	0.031	0.08
	3	Clinker//[RoW] clinker production	0.030	0.11
	4	Heat, district or industrial, other than natural gas//[RoW] heat production, at hard coal industrial furnace 1-10 MW	0.029	0.14
	5	Diesel, burned in building machine//[GLO] diesel, burned in building machine	0.024	0.17
	6	Pig iron//[GLO] pig iron production	0.023	0.19
	7	Electricity, high voltage//[IN] electricity production, hard coal	0.015	*0.21*
ReCiPe	1	Hard coal//[CN] hard coal mine operation	0.029	0.03
	2	Sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	0.029	0.06
	3	Petroleum//[RoW] petroleum and gas production, on-shore	0.023	0.08
	4	Electricity, high voltage//[CN] electricity production, hard coal	0.023	0.11
	5	Petroleum//[RME] petroleum production, onshore	0.023	0.13
	6	Hard coal//[RoW] hard coal mine operation	0.016	0.15
	7	Hard coal//[RNA] hard coal mine operation	0.014	*0.16*
Etox	1	Sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	0.212	0.21
	2	Scrap steel//[RoW] treatment of scrap steel, municipal incineration	0.142	0.36
	3	Scrap copper//[RoW] treatment of scrap copper, municipal incineration	0.137	0.50
	4	Spoil from hard coal mining//[GLO] treatment of spoil from hard coal mining, in surface landfill	0.076	0.57
	5	Spoil from lignite mining//[GLO] treatment of spoil from lignite mining, in surface landfill	0.045	0.62
	6	Slag, unalloyed electric arc furnace steel//[RoW] treatment of slag, unalloyed electric arc furnace steel, residual material landfill	0.026	0.65
	7	Natural gas, unprocessed, at extraction//[GLO] natural gas production, unprocessed, at extraction	0.015	*0.66*

processes. For example, "clinker//[RoW] clinker production" and "sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site" are consistently important across both ReCiPe and CC. At the same time, however, different LCIA indicators have, due to their different foci, quite different inventory support,[21] and thus prioritize different unit processes. For example, the third largest process of ReCiPe, "petroleum//[RoW] petroleum and gas production, on-shore," does not show up in the prioritization list of the other LCIA indicators. That is, the most important unit processes may correlate among a subset of LCIA indicators but show divergent results for others. It follows that a broad and diverse spectrum of LCIA indicators is needed to avoid a one-sided prioritization of database improvements.

[21] With inventory support, we refer to the total amount of unit processes in the database which have a contribution according to a particular LCIA indicator, i.e., which include an elementary flow addressing one of the characterization factors contained in a particular LCIA indicator (see Table S4 in SI1 (Electronic Supplementary Material) for some examples).

3.2 Prioritizing across the full set of LCIA indicators

3.2.1 Unit processes in the prioritization space

We use the algorithm elaborated in Section 2.2 in order to identify, for all 19 selected LCIA indicators, the minimum number of processes required to equal or to succeed a given ccv. Table 3 shows an extract for the causer perspective of the sorted ranking table (corresponding to Fig. 2, step 6). The full table and the table for the connector perspective are both shown in SI2 (Electronic Supplementary Material). Table 3 is sorted in the following way: first according to the process frequency in the first ccv, then according to the process frequency in the second ccv and so forth, until the last ccv is reached. This sorting procedure ensures that the order of processes reflects their actual relevance according to their process frequency in the ccv bins across all LCIA indicators.

According to this ranking procedure, the unit process "electricity, high voltage//[CN] electricity production, hard coal" turns out to be the most important unit process in the database as it has a process frequency of 10. This means that the process

Table 2 Excerpt of the summary table for the seven connectors with the largest mean contribution for three selected LCIA indicators. The cumulated mean of relative importance "$(\beta_{con})_{sort\_cum\_rel}$" and the cumulated mean of absolute importance "$(\beta_{con})_{sort\_cum}$" of the processes with the seven largest mean contributions are highlighted in italic format. The full list can be viewed in the SI2. Used geographical shortcuts: CN, China; RoW, Rest-of-the-World; GLO, global; RME, Middle East; RAS, Russia

LCIA	Rank	Name [product //[geographical location] activity]	$(\beta_{con})_{sort}$	$(\beta_{con})_{sort\_cum}$	$(\beta_{con})_{sort\_rel}$	$(\beta_{con})_{sort\_cum\_rel}$
CC	1	Electricity, high voltage//[CN] electricity production, hard coal	0.117	0.12	0.015	0.02
	2	Pig iron//[GLO] pig iron production	0.071	0.19	0.009	0.02
	3	Hard coal//[CN] hard coal mine operation	0.065	0.25	0.009	0.03
	4	Clinker//[RoW] clinker production	0.065	0.32	0.009	0.04
	5	Heat, district or industrial, other than natural gas//[RoW] heat production, at hard coal industrial furnace 1-10 MW	0.064	0.38	0.009	0.05
	6	Electricity, high voltage//[CN] market for electricity, high voltage	0.055	0.44	0.007	0.06
	7	Diesel, burned in building machine//[GLO] diesel, burned in building machine	0.055	*0.49*	0.007	*0.07*
ReCiPe	1	Petroleum//[GLO] market for petroleum	0.092	0.09	0.012	0.01
	2	Electricity, high voltage//[CN] electricity production, hard coal	0.065	0.16	0.008	0.02
	3	Hard coal//[CN] hard coal mine operation	0.063	0.22	0.008	0.03
	4	Sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	0.058	0.28	0.007	0.04
	5	Copper//[GLO] market for copper	0.049	0.33	0.006	0.04
	6	Petroleum//[RoW] petroleum and gas production, on-shore	0.048	0.37	0.006	0.05
	7	Petroleum//[RME] petroleum production, onshore	0.047	*0.42*	0.006	*0.05*
Etox	1	Sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	0.423	0.42	0.046	0.05
	2	Scrap steel//[RoW] treatment of scrap steel, municipal incineration	0.283	0.71	0.031	0.08
	3	Scrap copper//[RoW] treatment of scrap copper, municipal incineration	0.274	0.98	0.030	0.11
	4	Sulfidic tailing, off-site//[GLO] market for sulfidic tailing, off-site	0.211	1.19	0.023	0.13
	5	Spoil from hard coal mining//[GLO] treatment of spoil from hard coal mining, in surface landfill	0.151	1.34	0.016	0.14
	6	Scrap steel//[GLO] market for scrap steel	0.141	1.48	0.015	0.16
	7	Scrap copper//[GLO] market for scrap copper	0.137	*1.62*	0.015	*0.17*

contributes in such magnitude and across such a wide array of impact categories that its cumulated contribution is required to equal or succeed the ccv of 0.1 across 10 LCIA indicators. The ranking order of the 3 subsequent processes is not determined by the ccv of 0.1—they all contribute significantly, i.e., are located in the first ccv bin of 0.1, in regard to 3 LCIA indicators—but by their frequency in the succeeding ccv bins. Note that the processes prioritized by just one LCIA indicator in the ccv bin of 0.1 are often prioritized by another LCIA indicator in one of the higher bins. It is also noteworthy that part of the process frequency is not shown as it is beyond the ccv of 0.6.

As this format is difficult to interpret, we aggregate the process frequency of these processes to their corresponding ISIC sector and geographical location.

3.2.2 The most important sectors

Figure 4 shows the proportion of total process frequency per ccv differentiated into ISIC rev.4 sectors,[22] the activity-based classification system used by the ecoinvent database. For example, the first stacked bar (ccv of 0.1) shows the total process frequency (39) of the (22 largest) unit processes (shown in Table 3) per ISIC sector. Each bin illustrates only the additional process frequency and does not include the total process frequency of the preceding ccv.

Figure 4 shows that processes referring to electric power generation, treatment and disposal of hazardous waste, extraction of crude petroleum, and mining of hard coal have the largest overall effect on database quality (see first bar). Overall, electric power generation causes roughly 30% of the total process frequency. The connector perspective also highlights the frequency of the electricity sector for the transmission of environmental impacts (see Fig. S8, SI1, Electronic Supplementary Material). Processes related to electric power generation, manufacture of basic iron and steel, treatment and disposal of hazardous waste, extraction of crude petroleum, mining of hard coal, and freight transport are of relevance. They accumulate roughly 70% of the total process frequency in the first ccv. Surprisingly, market datasets constantly represent only around 40% of the connectors throughout the first ccv (see SI2, Electronic Supplementary Material).

Figure 4 also visualizes the sector relevance throughout the ccv's. From the total process frequency added throughout the

[22] International Standard Industrial Classification of All Economic Activities, Revision 4 (see http://unstats.un.org/unsd/cr/registry/isic-4.asp)

Table 3 The most relevant 25 unit processes (causer perspective) according to our set of LCIA indicators. The results are sorted incrementally according to their process frequency in the cumulated contribution bins. The full list is shown in SI2. Used geographical shortcuts: CN, China; RoW, Rest-of-the-World; GLO, global; RME, Middle East; RAS, Asia; CA-ON, Ontario; BR, Brazil; IN, India

Rank	Name [product //[geographical location] activity]	Total frequency across all LCIA indicators	Cumulated contribution value					
			0.1	0.2	0.3	0.4	0.5	0.6
1	Electricity, high voltage//[CN] electricity production, hard coal	14	10	0	0	0	0	0
2	Hard coal//[CN] hard coal mine operation	16	3	3	1	0	1	2
3	Petroleum//[RoW] petroleum and gas production, on-shore	10	3	1	0	1	0	0
4	Sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	8	3	0	0	2	1	0
5	Diesel, burned in building machine//[GLO] diesel, burned in building machine	12	2	5	1	0	0	0
6	Clinker//[RoW] clinker production	13	2	2	1	1	2	1
7	Heat, district or industrial, other than natural gas//[RoW] heat production, at hard coal industrial furnace 1-10 MW	13	1	4	3	1	0	0
8	Petroleum//[RME] petroleum production, onshore	12	1	3	0	2	0	0
9	Blasting//[RoW] blasting	9	1	2	1	2	0	0
10	Natural gas, high pressure//[RoW] natural gas production	14	1	1	2	1	1	1
11	Spoil from hard coal mining//[GLO] treatment of spoil from hard coal mining, in surface landfill	7	1	0	3	1	1	0
12	Slag, unalloyed electric arc furnace steel//[RoW] treatment of slag, unalloyed electric arc furnace steel, residual material landfill	8	1	0	1	0	0	0
13	Copper//[RAS] copper production, primary	12	1	0	0	1	2	2
14	Copper//[RoW] copper production, primary	12	1	0	0	1	1	2
15	Sawlog and veneer log, softwood, measured as solid wood under bark//[RoW] softwood forestry, pine, sustainable forest management	4	1	0	0	1	1	1
16	Electricity, high voltage//[CA-ON] electricity production, nuclear, pressure water reactor, heavy water moderated	6	1	0	0	0	1	1
17	Tailing, from uranium milling//[GLO] treatment of tailing, from uranium milling	5	1	0	0	0	0	2
18	Soybean//[BR] soybean production	14	1	0	0	0	0	0
19	High level radioactive waste for final repository//[RoW] treatment of high level radioactive waste for final repository	1	1	0	0	0	0	0
20	Water, decarbonised, at user//[RoW] water production and supply, decarbonised	2	1	0	0	0	0	0
21	Zinc concentrate//[GLO] zinc-lead mine operation	13	1	0	0	0	0	0
22	Oxygen, liquid//[RoW] air separation, cryogenic	2	1	0	0	0	0	0
23	Electricity, high voltage//[IN] electricity production, hard coal	15	0	2	2	4	2	0
24	Transport, freight, sea, transoceanic ship//[GLO] transport, freight, sea, transoceanic ship	12	0	2	2	3	1	0
25	Pig iron//[GLO] pig iron production	13	0	2	1	1	0	0

ccv's, between 12% (second ccv) and 32% (fifth ccv) refer to "electric power generation…" indicating its consistent relevance across all ccv's. The same insight applies to the extraction of crude petroleum and the mining of hard coal. The treatment and disposal of hazardous waste, in turn, is of particular relevance with regard to the first ccv. This results from the fact that the most important processes in this sector hold a large amount of contribution such as the mentioned "tailing, from uranium milling."

3.2.3 The most important locations

Figure 5 shows the proportion of total process frequency per ccv differentiated into geographical locations.

Figure 5 shows that the most important *causers* predominantly relate to geographical locations with a low spatial specificity, i.e., RoW, GLO, and CN. These locations are particularly relevant for the first ccv where they represent 90% of the process frequency, i.e., 36%, 33%, and 21% for RoW, CN, and GLO, respectively. The dominance of CN and GLO geographies decreases with increasing size of the ccv whereas RoW is consistently important across all ccv. The reason for this is that in many cases, local production datasets only cover a small proportion of the global production volumes. The relevance of RoW datasets therefore indicates a general lack of regionally appropriate data.

The *connector* perspective is also dominated by processes referring to RoW, GLO, and CN (see Fig. S9 in SI1, Electronic

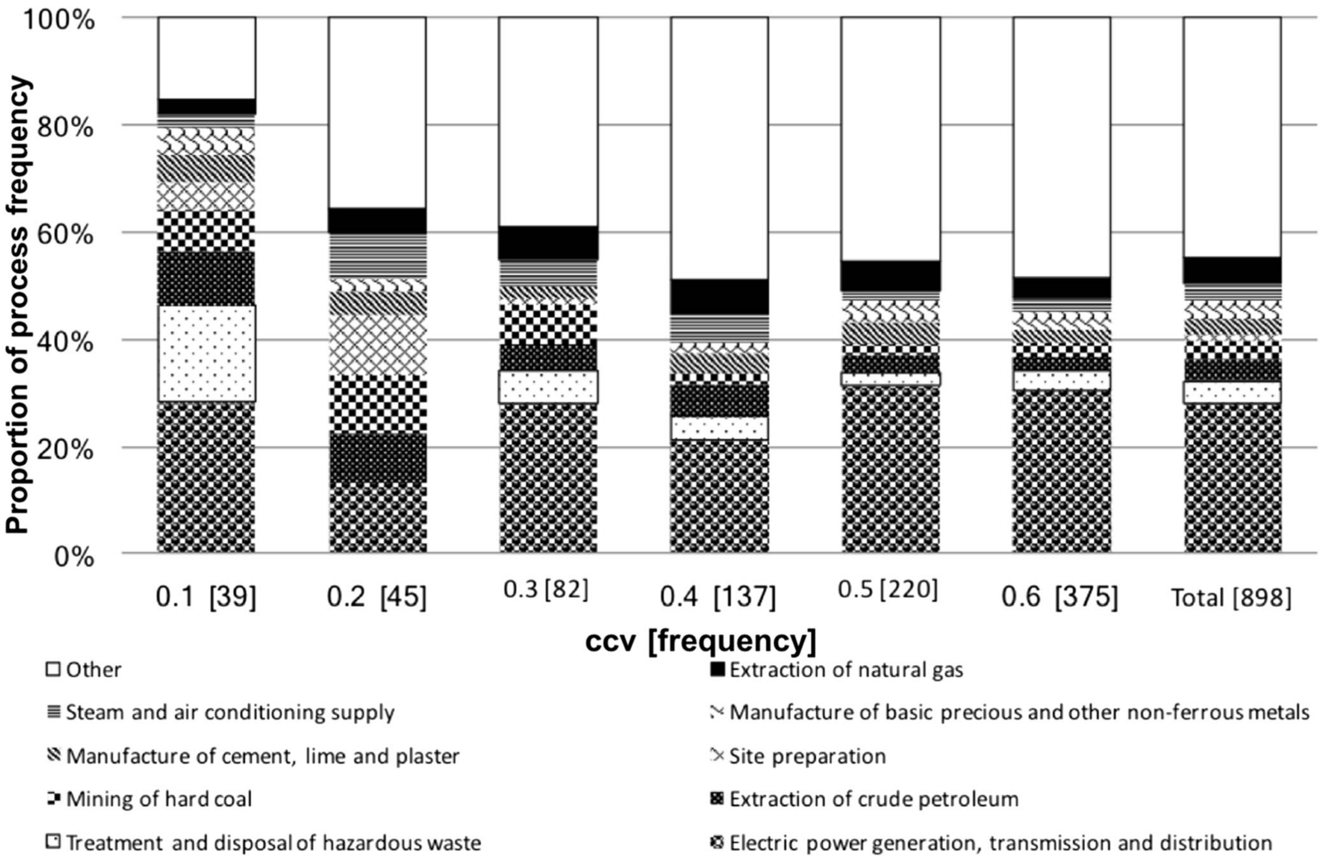


Fig. 4 Accumulated process frequency per ccv and ISIC rev. class. The total number of processes required to exceed a given ccv is given in square brackets. For clarity, only 9 ISIC categories are shown. The remaining 35 categories are consolidated into the "Other" category

Supplementary Material). Processes referring to these geographies cause almost 90% of the process frequency in the first ccv, i.e., 27%, 22%, and 40% for RoW, CN, and GLO, respectively. Overall, datasets belonging to these geographies generate about 62% of the total process frequency.

4 Discussion

The application of our prioritization approach to the ecoinvent database reveals a remarkable concentration in the distribution of internal process relevance, even when considering 19 selected LCIA indicators. That is, a relatively large proportion of the overall database quality is dependent on the quality of a small set of processes. Concentrating research efforts on the increase of information density in these systemic processes offer important starting points for the systematic and effective improvement of the entire database.

We investigated two prioritization perspectives which support the detection of two crucial characteristics of unit processes. The causer perspective prioritizes processes with exchanges of resources and emissions that are consistently important across product systems and LCIA indicators. Overall, 3% of the processes in the database cause more than 60% of the total process relevance. Processes referring to electricity generation, waste treatment activities, and energy carrier provision (petroleum and hard coal) are consistently important. The connector perspective prioritizes sensitive hubs whose modification can alter the results for the overall database considerably. In total, 8% of the processes transmit more than 50%, if normalized to the contribution transmitted by all connectors. In absolute terms, the contributions "flowing through" these most important connectors translate into roughly four times the total contribution in the database. This highlights the overall relevance of such hubs for database quality and suggests that they receive more attention. Particularly, electricity generation but also iron and steel production have strong "network effects."

The relatively low number of systemic datasets should not be automatically taken as a general characteristic of LCA or the economic system it strives to model. In fact, besides the high correlation among LCIA indicators already mentioned, it is also the evolution of partly arbitrary and partly intended modeling choices that explains the inequality in the database structure.

1. The presence of different *process types* that fosters inequality in relative process relevance. In addition to

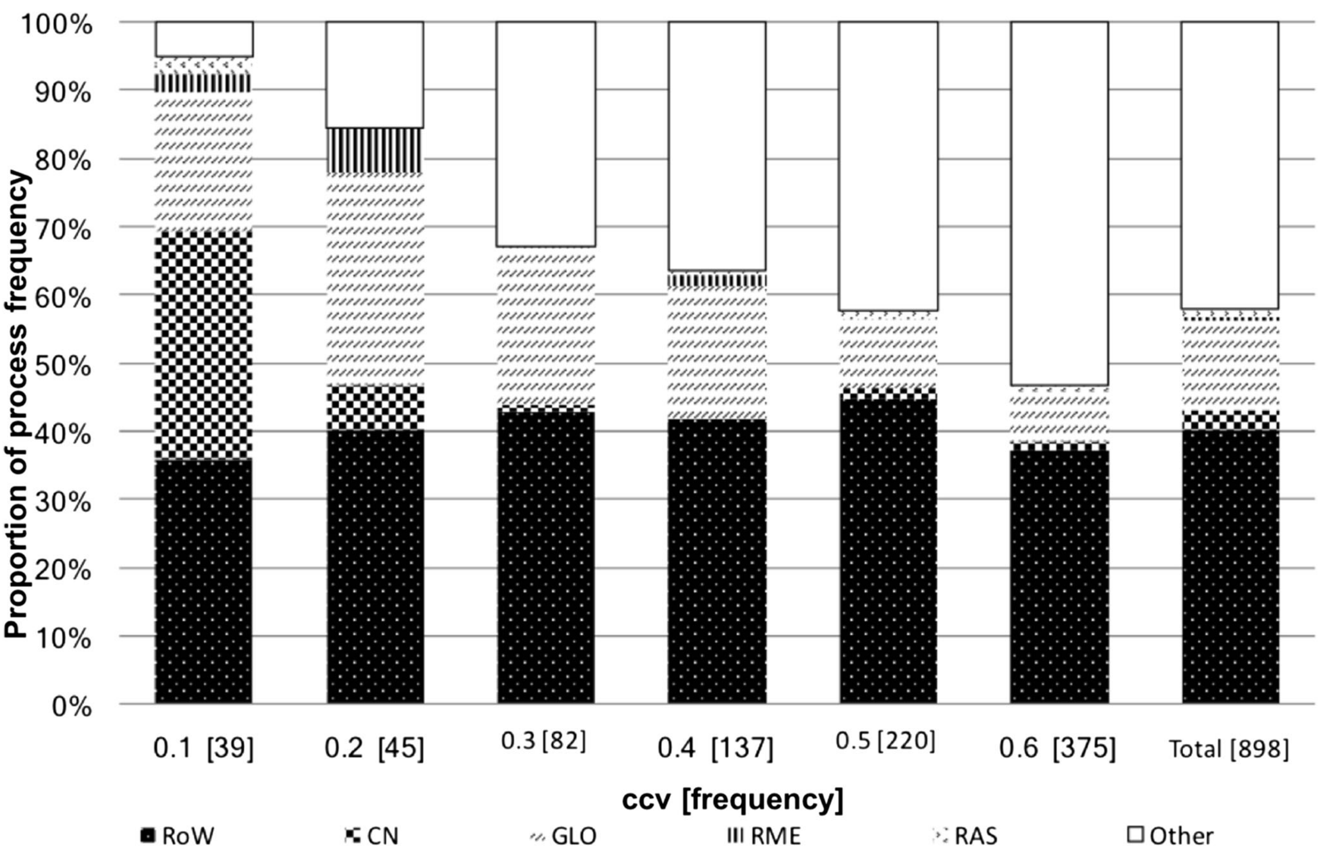


Fig. 5 Accumulated process frequency per ccv and geographical location. The total process frequency per ccv is given in square brackets. RoW, Rest-of-the-World; CN, China; GLO, global; RME, Europe; RAS, Russia. The remaining 44 geographical locations are consolidated under "Other"

transforming processes, ecoinvent has always used consolidating processes to represent geographical and technological averages or to maintain a certain modeling structure (e.g., production mixes or markets). Typically, such datasets cause no direct environmental intervention and consequently show no contribution in the causer perspective. Version 3.1 of the database maintains 3196 market datasets. That is, a large proportion of the processes in the database are *a priory* irrelevant for the causer perspective (see Table S3 in SI1, Electronic Supplementary Material).

2. Some elementary flows (resources or emissions) are only listed in a few unit processes. Consequently, the inventory support for some LCIA indicators is rather low (see Table S4 in SI1, Electronic Supplementary Material). This is not necessarily a characteristic of our physical reality but results, at least in part, from the fact that more attention has been paid to some emissions or impact categories than others. For example, among the assessed mid-point methods, LCIA indicators related to climate change have the largest inventory support—3500 processes cause emission with a global warming potential—while agricultural land occupation and resource depletion indicators appear at the other end of the scale with an inventory support of roughly 300 processes. The endpoint method with the largest inventory support is the Swiss ecological scarcity method (2013) with roughly 6000 processes (see Figs. S2 and S4 in SI1, Electronic Supplementary Material). The inventory support seems to be one important driver for the inequality observed among different LCIA indicators, i.e., the degree of concentration in β_{cau} and β_{con} as measured by the Gini coefficient (see Table S4 in SI1, Electronic Supplementary Material).
3. The ecoinvent database has been evolving as an incomplete model of our economic system. The incomplete spatial and technological coverage has at least two notable effects. First, it limits the selection of appropriate intermediate inputs required for the accurate representation of new processes. This fosters, at least to a certain degree, the iterative use of the same (generic) intermediate inputs and favors a network structure with consistent dependencies. For example, the relevance of the process "diesel, burned in building machine//[GLO]" (rank 5) results, at least in part, from its lack in spatial and technological specificity and the lack of alternatives. Second, the incompleteness in spatial process support causes (since version 3.0) the repetitive linking to generic geographical

locations and consequently promotes the relevance of such locations. Whenever local datasets are unavailable for the supply of a certain product, the product is supplied by the GLO dataset. If local production datasets are available but only cover a small proportion of the global production volumes, most of the production volume is supplied by a RoW dataset. That is, the overrepresentation of generic geographical locations in the first ccv indicates a clear lack in spatial process support.

Some issues of importance need to be considered with regard to the prioritization method. First, our approach represents an internal perspective from which the optimal allocation of inventory efforts follows relative process relevance across different LCIA indicators. This is useful for the identification of consistently important processes but provides no direct[23] support for the identification of "blank spots" and is therefore "unable to direct research efforts to economic sectors that may be underrepresented" (Reinhard et al. 2016). Therefore, improvement efforts around LCI databases should not be exclusively guided by internal process relevance. In fact, the proposed ranking of most relevant processes should be seen as a complement to already existing routines and approaches. Its inward-oriented perspective should be complemented with more outward-oriented prioritization methods such as the one presented by Majeau-Bettez et al. (2011).

Further, our ranking algorithm ensures, for a given set of LCIA indicators, that all processes required to equal or succeed a certain ccv are considered. It operates on the basis of the Lorenz curve which is a common tool for analyzing and comparing income inequality (Duclos and Araar 2006). However, the discretization to size classes comes at the cost of information loss, notably large contributions are reduced to the size class of the ccv. In general, as the order of processes is first determined by their actual size class (the ccv) and only then by their process frequency, the sorting procedure preserves essential information about the actual size of a unit process contribution. It ensures that processes in the first ccv, only prioritized by one rather uncorrelated LCIA indicator (such as ALOP or IR), will still receive more attention than a process which is relevant according to many LCIA indicators in the second ccv. We recognize the potential utility of this ranking procedure within every LCA application which involves the use of many LCIA indicators (Hellweg and Milà i Canals 2014). That is, the algorithm could be used to identify key process datasets in every standard LCA application. The selection of the ccv in this article is rather coarse; the complete prioritization list presented in SI2 (Electronic Supplementary Material) uses a much finer scale. Note, however, that the amount of the ccv's only affects the ranking of the processes within a particular ccv but not the actual amount of processes associated with this ccv. That is, the same 22 unit processes shown in Table 3 are required to equal or exceed the ccv of 10%, independent on the amount of used ccv's.

The discretization also guarantees a basic equivalence among the LCIA indicators. While this avoids (implicit) weighting of LCIA indicators on the basis of the inequality in their mean of absolute importance, it means that each LCIA indicator is treated equally. That is, a unit process added by the midpoint indicator "climate change" (IPCC2007) is considered of equal importance than a unit process added by the endpoint indicator "ReCiPe total." One can imagine multi-criteria approaches which assign different weights to different LCIA indicators. However, as we keep all unique unit processes on the prioritization list, such a weighting would only affect the rank of the processes *in* the list but not change the actual amount of processes *on* the list.

The actual amount of processes *on* the list is dependent on the set of selected LCIA indicators. Each LCIA indicator can be considered as a unique optimization vector. Therefore, adding or removing LCIA indicators from the selected set can change the amount of prioritized processes. Our prioritization method is applicable to any set of LCIA indicators and future work should analyze different LCIA indicator perspectives, e.g., a set of midpoint methods versus a set of endpoint methods, and the corresponding differences in their recommendations. This would reveal the detailed inventory support of different LCIA paradigms and therefore offer important feedback for LCIA method and LCI database developers.

5 Conclusions

Knowing the unit processes with the largest contribution is a prerequisite of systematic, targeted LCI database improvement. We demonstrated a contribution-based prioritization approach using version 3.1 of the ecoinvent database as a practically relevant example. Our approach facilitates the alignment of data collection efforts to unit processes with the highest relevance in terms of their overall influence on the results calculated with many different LCIA indicators. Focusing research efforts on these processes, which play a dominant role in nearly every LCA application, allows for effective improvement of the datasets.

The list of most important unit processes we presented offers new starting points for the effective improvement of the ecoinvent database as it allows the allocation of available resources according to the structural dependencies in the data. This strengthens the basis for decision-making of the people managing the data, who can systematically process the list and

[23] Note, however, that the method offers some indirect support for the identification of blank spots, namely via the identification of processes which are modeled at a too generic level. Such process models should be replaced by more specific ones in a spatial or in technological scale.

decide for each unit process if and how it should be improved. To improve LCI database quality where it matters the most, unit processes of systemic relevance should achieve a high data quality rating.

Considering the high cost of data collection, the improvement of LCI databases should be organized as effectively as possible. We observe a general lack of operational tools for the analysis of LCI databases which help to realign the focus of inventory efforts to more systematic choices—as opposed to choices guided by external data availability and situation-driven requirements. Certainly, such data and requirements have a great importance, but future research should also ensure that coverage and specificity of LCI databases progress towards a more representative model of the complex economic system LCA strives to capture at its physical level. The structural dependencies in the example database, as discussed in this article, represent a meaningful perspective for improvement but should be complemented with existing routines and other, more outward-looking prioritization approaches.

Acknowledgements We gratefully acknowledge Christopher L. Mutel for the valuable comments and suggestions given to improve the manuscript. We are thankful to the four anonymous reviewers for their insightful comments and inputs.

References

Bourgault G, Lesage P, Samson R (2012) Systematic disaggregation: a hybrid LCI computation algorithm enhancing interpretation phase in LCA. Int J Life Cycle Assess 17:774–786

Ciroth A, Muller S, Weidema B, Lesage P (2016) Empirically based uncertainty factors for the pedigree matrix in ecoinvent. Int J Life Cycle Assess 21:1338–1348

Duclos J-Y, Araar A (2006) Poverty and equity: measurement, policy, and estimation with DAD. Springer

Ebner F (2013) Results, interpretation and reporting: GaBi Software. In: Gabi-Softw. http://www.gabi-software.com/software/gabi-software/gabi-5/functionalities/results-interpretation-and-reporting/. Accessed 16 Aug 2013

EC-JRC (2011) International Reference Life Cycle Data System (ILCD) Handbook- Recommendations for Life Cycle Impact Assessment in the European context. First edition November 2011. EUR 24571 EN. Luxemburg. Publications Office of the European Union

Finnveden G, Hauschild MZ, Ekvall T, Guinée J, Heijungs R, Hellweg S, Koehler A, Pennington D, Suh S (2009) Recent developments in Life Cycle Assessment. J Environ Manag 91:1–21

Goedkoop M, Oele M (2004) SimaPro 6 - Introduction to LCA with SimaPro

Heijungs R, Huijbregts M (2004) A review of approaches to treat uncertainty in LCA. Complex Integr Resour Manag Proc 2nd Bienn Meet Int Env Model Softw Soc

Heijungs R, Kleijn R (2001) Numerical approaches towards life cycle interpretation five examples. Int J Life Cycle Assess 6:141–148

Heijungs R, Suh S (2002) The computational structure of life cycle assessment. Kluwer Academic Publishers, Dordrecht

Hellweg S, Milà i Canals L (2014) Emerging approaches, challenges and opportunities in life cycle assessment. Science 344:1109–1113

Huijbregts MJ (1998) Application of uncertainty and variability in LCA. Int J Life Cycle Assess 3:273–280

Majeau-Bettez G, Strømman AH, Hertwich EG (2011) Evaluation of process- and input–output-based life cycle inventory data with regard to truncation and aggregation issues. Environ Sci Technol 45: 10170–10177

Mutel C (2012) Framework and tools for regionlization in life cycle assessment. ETH

Reinhard J, Mutel C, Wernet G et al (2016) Contribution-based prioritization of LCI database improvements: method design, demonstration, and evaluation. Environ Model Softw 86:204–218

Sonnemann G, Vigon B (eds) (2011) Global guidance principles for life cycle assessment databases: a basis for greener processes and products (Shonan Guidance Principles)

Steubing B, Wernet G, Reinhard J, Bauer C, Moreno-Ruiz E (2016) The ecoinvent database version 3 (part II): analyzing LCA results and comparison to version 2. Int J Life Cycle Assess 21:1269–1281

Tillman A-M (2000) Significance of decision-making for LCA methodology. Environ Impact Assess Rev 20:113–123

Weidema B, Bauer C, Hischier R et al (2013) Overview and methodology. Data quality guideline for the ecoinvent database version 3

Wernet G, Bauer C, Steubing B, Reinhard J, Moreno-Ruiz E, Weidema B (2016) The ecoinvent database version 3 (part I): overview and methodology. Int J Life Cycle Assess 21:1218–1230